ランキング関数のオンライン学習について (計算機科学基礎理論とその応用)

AUTHOR(S):

中村, 篤祥

CITATION:

中村, 篤祥. ランキング関数のオンライン学習について (計算機科学基礎理論とその応用). 数理解析研究所講究録 2005, 1426: 51-56

ISSUE DATE:

2005-04

URL:

http://hdl.handle.net/2433/47298

数理解析研究所講究録 1426 巻 2005 年 51-56

ランキング関数のオンライン学習について

北海道大学大学院情報科学研究科 中村篤祥 (Atsuyoshi Nakamura)
Graduate School of Information Science and Technology, Hokkaido University

1 はじめに

ランキングとは、多クラス分類の1種で、クラス間に全順序の関係がなりたつ場合をいう。本論文では、ランキング関数として、法線ベクトルとクラス境界の閾値により表現される線形ランキング関数について考える。線形ランキング関数に関し、Rajaram ら [7] は VC 次元の解析を行い、線形識別子と差が無いことを示した。ここではグラフ次元の解析を行い、線形ランキング関数が線形識別子からなる決定リストと VC 次元は同じであるが、グラフ次元においてはより複雑である可能性を示す。また、Crammer と Singer による Perceptron アルゴリズムを拡張した Prank アルゴリズムについて、彼らが証明したオンライン学習におけるランキング損失の上界に関する定理を、マージンが負の場合にも拡張する。

2 線形ランキング関数

X を集合、$Y=\{1,2,...,k\}$ とする。このとき、$f:X\to Y$ を **k 値関数**と呼ぶ。本論文では $X=\Re^n$ とし、k を固定して考える。$\mathbf{w}\in\Re^n$ と $b_1\leq b_2\leq\cdots\leq b_{k-1}$ を満たす $\mathbf{b}=(b_1,b_2,...,b_{k-1})$ により以下のように定義される f を**線形ランキング関数**と呼ぶ。

$$f(\mathbf{x})=\min_{r\in Y}\{r:\mathbf{w}\cdot\mathbf{x}-b_r<0\}$$

ただし、$b_k=\infty$ とする。

線形ランキング関数の族を $\mathcal{R}$ で表す。

3 線形ランキング関数の複雑さ

関数族の複雑さを比較するために、線形ランキング関数族を含む、より大きな関数族を考える。いま、$\{-1,1\}$-値の関数族 $\mathcal{B}$ に属する $k-1$ 個の関数 $g_1,g_2,...,g_{k-1}$ により定義される、決定リストの形をした次のような k 値関数 f を考える。

$$f=[(g_1,1),(g_2,2),...,(g_{k-1},k-1),(g_k,k)]$$

ただし、$g_k=1$ とし、f は $f(x)=\min\{i:g_i(x)=1\}$ で定義される関数を表しているものとする。このように定義される関数 f からなる関数族を $\mathcal{F}_{\mathcal{B}}$ とする。$\mathcal{L}$ を線形識別関数族とすれば、$\mathcal{F}_{\mathcal{L}}$ は線形ランキング関数族 $\mathcal{R}$ を含む。この節では、これら2つの関数族の複雑さを分析し比較する。

3.1 VC 次元による分析

$\mathcal{F}$ を k 値関数の集合とする。任意の自然数 l に対し、X の l 個要素からなるリスト $S=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_l)\in X^l$ を考える。$f\in\mathcal{F}$ に対して、$f_S=(f(\mathbf{x}_1),f(\mathbf{x}_2),...,f(\mathbf{x}_l))\in Y^l$ とする。このとき、

$$\Pi_{\mathcal{F}}(S)=\{f_S:f\in\mathcal{F}\}$$

と定義する。更に、自然数 m に対して

$$\Pi_{\mathcal{F}}(m) = \max_{S \in X^m} |\Pi_{\mathcal{F}}(S)|$$

と定義する。ただし、$|\cdot|$ は集合の要素数を表すものとする。$|\Pi_{\mathcal{F}}(S)| = k^m$ のとき、S は $\mathcal{F}$ により shatter されるという。$\mathcal{F}$ により shatter される最大要素数の集合の要素数を $\mathcal{F}$ の VC 次元という。つまり、k 値関数族 $\mathcal{F}$ の VC 次元 $d_V(\mathcal{F})$ は、

$$d_V(\mathcal{F}) = \max\{m : |\Pi_{\mathcal{F}}(m)| = k^m\}$$

と定義される [1, 7][1]。

次の命題が成り立つことは明らかである。

命題 1 $d_V(\mathcal{F}_{\mathcal{B}}) \leq d_V(\mathcal{B})$

命題 1 と $\mathcal{R} \subseteq \mathcal{F}_{\mathcal{L}}$ より

$$d_V(\mathcal{R}) \leq d_V(\mathcal{L}) \tag{1}$$

が成り立つ。[7] では、不等式 (1) において等号が成立することが示されている。

定理 1 (Rajaram et al. [7]) $d_V(\mathcal{R}) = d_V(\mathcal{L})$

したがって、$d_V(\mathcal{R}) = d_V(\mathcal{F}_{\mathcal{L}})$ となり、VC 次元による複雑さの分析では、2 つの関数族に差がでないことがわかる。

3.2 グラフ次元による分析

2 値関数族の VC 次元の k 値関数族への拡張として、Natarajan は以下のように定義されるグラフ次元というものを考えた [5, 1].

まず、$Y \times Y$ 上の $\{0,1\}$-値関数 δ を、$i = j$ のときのみ $\delta(i,j) = 1$ となる関数とする。任意の自然数 l に対し、X の l 個要素からなるリスト $S = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l) \in X^l$ を考える。$I = (i_1, i_2, ..., i_l) \in Y^l$ 及び $f \in \mathcal{F}$ に対して、

$$f_{I,S} = (\delta(f(\mathbf{x}_1), i_1), \delta(f(\mathbf{x}_2), i_2), ..., \delta(f(\mathbf{x}_l), i_l)) \in \{0,1\}^l$$

とする。このとき、自然数 m に対して

$$\begin{aligned} \Pi_{I,\mathcal{F}}(S) &= \{f_{I,S} : f \in \mathcal{F}\} \\ \Pi_{\mathcal{F}}(m) &= \max_{S \in X^m, I \in Y^m} |\Pi_{I,\mathcal{F}}(S)| \end{aligned}$$

と定義する。

k 値関数族 $\mathcal{F}$ のグラフ次元 $d_G(\mathcal{F})$ は、

$$d_G(\mathcal{F}) = \max\{m : |\Pi_{\mathcal{F}}(m)| = 2^m\}$$

と定義される。

命題 2 $d_G(\mathcal{F}_{\mathcal{B}}) \leq k d_V(\mathcal{B})$

[1] $d_V(\mathcal{F})$ のことを、論文 [7] では**ランキング次元**と呼んでいる。

(証明) $I=(i_1,i_2,...,i_l)$ において $I_i=\{j:i_j=i\}$ が $d_V(\mathcal{B})$ 個より多くあるとする。任意の $S=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_l)$ に対し、$S_i=\{\mathbf{x}_j:j\in I_i\}$ とする。このとき、S_i のある部分集合 S が存在し、$\mathcal{B}$ に属する関数で S と S_i-S を分離するものが存在しない。このとき、これをランク境界としてもつ $\mathcal{F}_\mathcal{B}$ の関数は存在しない。したがって $d_G(\mathcal{F}_\mathcal{B})\leq kd_V(\mathcal{B})$ が成り立つ。 □

命題 3 $(k-1)n\leq d_G(\mathcal{F}_\mathcal{L})\leq k(n+1)$

(証明) 命題 2 と線形関数の VC 次元が $n+1$ であるという事実から、$d_G(\mathcal{F}_\mathcal{L})\leq k(n+1)$ が成り立つ。

以下、$kn\geq d_G(\mathcal{F}_\mathcal{L})$ を証明する。空間 $X=\{(x_1,x_2,...,x_n):x_i\in\Re\}$ において、$x_1=0$ という部分空間を考える。これは、空間 $\Re^{n-1}$ とみなせるので、この空間における線形識別関数の VC 次元は n である。そこで線形識別関数により shatter される n 点のリスト S_0 を考える。S_0 を x_1 軸方向に i だけ移動した集合を S_i とする。$S_0,S_1,...,S_{k-2}$ をつなげてできる長さ $(k-1)n$ のリスト S と

$$I=(\underbrace{1,1,...,1}_{n\text{ times}},\underbrace{2,2,...,2}_{n\text{ times}},...,\underbrace{k-1,k-1,...,k-1}_{n\text{ times}})$$

を考えたとき $\Pi_{I,\mathcal{F}}(S)=\{0,1\}^{(k-1)n}$ であることを示す。任意の $A\in\{0,1\}^{(k-1)n}$ が与えられたとする。S_i を含む超平面 $x_1=i$ 内において、線形関数 g_i が存在し、S_i に含まれる n 個の点に対し、対応する A 内の要素の値が 1 のときのみ g_i の値が負となる。この線形関数は、全空間の線形関数 f_i を $x_1=i$ 内に制限したものとみることができる。関数 f_i の法線方向は $(1,0,0,...,0)$ にいくらでも近くとれる。つまり、S_{i+1} 内の全ての点 $\mathbf{x}$ に対し、$f_i(\mathbf{x})>0$ とできる。このとき、$f=[(f_1,1),(f_2,2),...,(f_{k-1},k-1),(1,k)]$ に対し、$f_{I,S}=A$ が成り立つ。 □

命題 4 $d_G(\mathcal{R})\geq n+k-1$

(証明) 命題 3 と同様に、空間 $X=\{(x_1,x_2,...,x_n):x_i\in\Re\}$ において、$x_1=0$ という空間内の点で、線形識別関数により shatter される n 点のリスト S_0 を考える。また、$i=1,2,...,k-1$ に対して 1 点からなるリスト $S_i=\{(i,0,0,...,0)\}$ を考える。$S_0,S_1,...,S_{k-1}$ をつなげてできる長さ $n+k-1$ のリスト $S=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_{n+k-1})$ と

$$I=(\underbrace{1,1,...,1}_{n\text{ times}},2,3,...,k)$$

を考えたとき $\Pi_{I,\mathcal{F}}(S)=\{0,1\}^{n+k-1}$ であることを示す。任意の $A=(a_1,a_2,...,a_{n+k-1})\in\{0,1\}^{n+k-1}$ が与えられたとする。命題 3 と同様な議論により、ある $\mathbf{w}\in\Re^n$ と $b_1\leq b_2\leq...\leq b_{k-1}$ が存在し、これにより定義される線形ランキング関数 $f(\mathbf{x})=\min_{r\in Y}\{r:\mathbf{w}\cdot\mathbf{x}-b_r<0\}$ により、S_0 内においては、対応する A 内の要素が 1 のときのみ $f(\mathbf{x})=1$ となり、$S_i=\{\mathbf{x}_{n+i}\}(i>0)$ においては、$f(\mathbf{x}_{n+i})=i+1$ となる。したがって、$(a_{n+1},a_{n+2},...,a_{n+k-1})=(1,1,...,1)$ の場合、$f_{I,S}=A$ が成り立つ。$(a_{n+1},a_{n+2},...,a_{n+k-1})\neq(0,0,...,0)$ の場合、関数 f の閾値 $b_2,b_3,...,b_{k-1}$ を次のような $b'_2,b'_3,...,b'_{k-1}$ に変更することにより $f_{I,S}=A$ が成り立つ。

$$b'_i=\begin{cases}b_i & (a_{n+i-1}=1)\\ b_{i-1} & (a_{n+i-1}=0,i<\max\{j:a_{n+j-1}=1\})\\ b_{i+1} & (a_{n+i-1}=0,i>\max\{j:a_{n+j-1}=1\})\end{cases}$$

$(a_{n+1},a_{n+2},...,a_{n+k-1})=(0,0,...,0)$ の場合、関数 f の $\mathbf{w}$ を $x_1=0$ の面に対称な方向にし、閾値 $b_2,b_3,...,b_{k-1}$ をすべて b_1 にすることにより $f_{I,S}=A$ が成り立つ。 □

残念ながら、$d_G(\mathcal{R})$ の明らかでない上界は今のところ求まっていないが、パラメータの数から考えると $d_G(\mathcal{R})\leq n+k-1$ ではないかと予想される。

4 無矛盾仮説出力アルゴリズムのPAC学習サンプル数

$\mathcal{D}$ を $X \times Y$ 上の確率分布とする。$h : X \to Y$ に対し、

$$\Pr_{(x,y)\sim\mathcal{D}}(h(\mathbf{x}) \neq y) < \epsilon$$

が成り立つとき、h を ϵ 近似仮説であるという。

グラフ次元に関し、Ben-David らは次のような定理を導いた。

定理 2 (Ben-David et al. [1]) $\mathcal{F}$ を X 上の k 値関数族とする。$\mathcal{D}$ を $X \times Y$ 上の確率分布で、すべての $(\mathbf{x}, i) \in X \times Y$ に対して、$\mathcal{D}(i|\mathbf{x}) = 1$ または 0 が成り立つものとする。このとき、ある定数 $c > 0$ が存在し、任意の $0 < \epsilon, \delta < 1$ に対し、確率分布 $\mathcal{D}$ にしたがってランダムに生成された

$$m \geq \frac{c}{\epsilon}\left(d_G(\mathcal{F}) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right) \tag{2}$$

個の例に無矛盾な $\mathcal{F}$ に属する仮説が ϵ 近似仮説である確率は、$1 - \delta$ より大きい。

したがって、$d_G(\mathcal{F}_\mathcal{L}) = \Theta(nk)$ であるから、$\mathcal{F}_\mathcal{L}$ の関数を学習する場合、k と n に関し $O(kn)$ でサンプル数が増加する。それに対し $d_G(\mathcal{R}) = O(n + k)$ の予想が正しければ、$\mathcal{R}$ に関してサンプル数の増加のオーダーを $O(n + k)$ に抑えることができる。

5 Prankアルゴリズムのマージンベースの性能評価

例 $(\mathbf{x}, y) \in X \times Y$ の線形ランキング関数 $f(\mathbf{x}) = \min_{r\in Y}\{r : \mathbf{w}\cdot\mathbf{x} - b_r < 0\}$ に対する**マージン** $\gamma((\mathbf{x}, y), f)$ を以下のように定める。

$$\gamma((\mathbf{x}, y), f) \stackrel{def}{=} \min\{(\mathbf{w}\cdot\mathbf{x} - b_r)y_r : r = 1, 2, ..., k-1\}$$

ただし、

$$(y_1, y_2, ..., y_{k-1}) = (\underbrace{1, 1, ..., 1}_{(y-1)times}, -1, -1, ..., -1)$$

とする。また、サンプル $S \subseteq X \times Y$ の線形ランキング関数 f に対するマージン $m(S, f)$ を

$$m(S, f) \stackrel{def}{=} \min_{(\mathbf{x},y)\in S} \gamma((\mathbf{x}, y), f)$$

で定義する。

Perceptron アルゴリズム [4] は、線形識別関数をオンライン学習するアルゴリズムである。Crammer と Singer[3] は、Perceptron アルゴリズムの線形ランキング関数版である Prank アルゴリズム (図 1) を考えた。

予測値 $\hat{y}$ と実際の値 y との差 $|\hat{y} - y|$ を**ランキング損失**という。このとき、Crammer と Singer[3] は以下の定理が成り立つことを示した。

定理 3 (Crammer and Singer [3]) 例のシークエンス $(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), ..., (\mathbf{x}^T, y^T)$ が、この順で与えられるとする。$||(\mathbf{w}^*, b_1^*, b_2^*, ..., b_{k-1}^*)|| = 1$ を満たすある線形ランキング関数 $f(\mathbf{x}) = \min_{r\in Y}\{r : \mathbf{w}^*\cdot\mathbf{x} - b_r^* < 0\}$ が存在し、このシークエンスに含まれる例の集合 S の f に対するマージンが $\gamma > 0$ であるとき、*Prank* アルゴリズムの累積ランキング損失は、高々 $(k-1)(R^2+1)/\gamma^2$ である。ただし、$R^2 = \max_t ||\mathbf{x}^t||^2 = \max_t \sum_{i=1}^n (x_i^t)^2$ とする。

初期値：$\mathbf{w}=\mathbf{0}, \mathbf{b}=(b_1, b_2, ..., b_{k-1})=(0, 0, ..., 0)$
以下の手順を繰り返す。

1. $\mathbf{x} \in X$ が与えられる。
2. 予測値 $\hat{y}=\min_{r \in Y}\{r : \mathbf{w} \cdot \mathbf{x}-b_r<0\}$ を計算する。
3. 実際の値 y が与えられる。
4. 以下のように $\mathbf{w}$ と $\mathbf{b}$ の更新を行う。
   (a) $\tau=(\tau_1, \tau_2, ..., \tau_{k-1})$ を以下のように定める。

$$\tau_r=\begin{cases} 1 & (\hat{y} \leq r<y) \\ -1 & (y \leq r<\hat{y}) \\ 0 & (otherwize) \end{cases}$$

   (b) $\mathbf{w}$ と $\mathbf{b}$ を次のように更新する。

$$\mathbf{w} \leftarrow \mathbf{w}+(\sum_{r=1}^{k-1}\tau_r)\mathbf{x}$$
$$\mathbf{b} \leftarrow \mathbf{b}-\tau$$

図 1: Prank アルゴリズム

定理 3 は、Perceptron アルゴリズムに関する Novikoff の定理 [6] の拡張になっている。この定理を拡張することにより、与えられた例の集合 S に関して、$m(S, f)<0$ となる線形ランキング関数 f に対する累積ランキング損失の上界も導ける。

定理 4 例のシークエンス $(\mathbf{x}^1, y^1), ..., (\mathbf{x}^T, y^T)$ が、この順で与えられるとする。$||(\mathbf{w}^*, b_1^*, b_2^*, ..., b_{k-1}^*)||=1$ を満たすある線形ランキング関数 $f(\mathbf{x})=\min_{r \in Y}\{r : \mathbf{w}^* \cdot \mathbf{x}-b_r^*<0\}$ に対し、このシークエンスに含まれる例の集合を S とし、y^t を $f(\mathbf{x}^t)$ に変更した集合 $S'=\{(\mathbf{x}^1, f(\mathbf{x}^1)), ..., (\mathbf{x}^T, f(\mathbf{x}^T))\}$ を考える。S を f で予測した場合の累積ランキング損失 $\sum_{t=1}^{T}|f(\mathbf{x}^t)-y^t|$ を L とする。S' に対する f のマージンが $\gamma>0$ であるとき[2]、*Prank* アルゴリズムの累積ランキング損失は、高々 $(k-1)(R^2+1)/\gamma^2+2L\left(1+\sqrt{R^2+1}/\gamma\right)$ である。ただし、$R^2=\max_t||\mathbf{x}^t||^2=\max_t\sum_{i=1}^{n}(x_i^t)^2$ とする。

(証明) $\mathbf{b}^*=(b_1^*, b_2^*, ..., b_{k-1}^*)$ とする。$\mathbf{v}^*=(\mathbf{w}^*, \mathbf{b}^*)$ とおく。また、Prank アルゴリズムにおいて、例 $(\mathbf{x}^t, y^t)$ に対する処理を行うときの更新前の $\mathbf{v}=(\mathbf{w}, \mathbf{b})$ 及び τ を、それぞれ $\mathbf{v}^t=(\mathbf{w}^t, \mathbf{b}^t)$ 及び τ^t とおく。このとき、

$$\mathbf{v}^* \cdot \mathbf{v}^{t+1}=\mathbf{v}^* \cdot \mathbf{v}^t+\sum_{r=1}^{k-1}\tau_r^t(\mathbf{w}^* \cdot \mathbf{x}^t-b_r^*) \tag{3}$$

が成り立つ。

$$\mathbf{y}^t=(y_1^t, ..., y_{k-1}^t)=(\underbrace{1, ..., 1}_{(y^t-1)times}, -1, ..., -1),\quad \mathbf{y}^*=(y_1^*, ..., y_{k-1}^*)=(\underbrace{1, ..., 1}_{(f(\mathbf{x}^t)-1)times}, -1, ..., -1)$$

とする。$(\mathbf{x}^t, y^t)$ に対する f のランキング損失を h^t とおくと、2つのベクトル $\mathbf{y}^t$ と $\mathbf{y}^*$ で値が異なる要素の数は h^t である。式 (3) の右辺の第2項に関して、$\tau_r^t \neq 0$ であれば $y_r^t=\tau_r^t$ であり、$y_r^*(\mathbf{w}^* \cdot \mathbf{x}^t-b_r^*) \geq \gamma$ であるので、例 $(\mathbf{x}^t, y^t)$ に対する Prank アルゴリズムのランキング損失を n^t とすれば、

$$\sum_{r=1}^{k-1}\tau_r^t(\mathbf{w}^* \cdot \mathbf{x}^t-b_r^*) = \sum_{\tau_r^t \neq 0, y_r^t=y_r^*} y_r^*(\mathbf{w}^* \cdot \mathbf{x}^t-b_r^*)+\sum_{\tau_r^t \neq 0, y_r^t \neq y_r^*} y_r^t(\mathbf{w}^* \cdot \mathbf{x}^t-b_r^*)$$

[2] S' の定義より $\gamma \geq 0$ は常に成り立つ。

$$
\begin{aligned}
&\geq (n_t - h_t)\gamma - \sum_{\tau_r^t \neq 0, y_r^t \neq y_r^*} |\mathbf{w}^* \cdot \mathbf{x}^t - b_r^*| \\
&= (n_t - h_t)\gamma - \sum_{\tau_r^t \neq 0, y_r^t \neq y_r^*} |\mathbf{v}^* \cdot (\mathbf{x}^t, -\mathbf{1}_r)| \\
&\geq (n_t - h_t)\gamma - \sum_{\tau_r^t \neq 0, y_r^t \neq y_r^*} ||\mathbf{v}^*||||(\mathbf{x}^t, -\mathbf{1}_r)|| \\
&\geq (n^t - h^t)\gamma - h_t\sqrt{R^2+1} \qquad (4)
\end{aligned}
$$

ただし、$\mathbf{1}_r$ は第 r 成分のみ 1 でその他は 0 の $(k-1)$ 次元ベクトルを表すものとする。

式 (3) と式 (4) より、

$$
\mathbf{v}^* \cdot \mathbf{v}^{T+1} \geq \gamma \sum_{t=1}^{T} (n^t - h^t) - \sqrt{R^2+1} \sum_{t=1}^{T} h_t = \gamma \sum_{t=1}^{T} n_t - L\left(\gamma + \sqrt{R^2+1}\right)
$$

よって

$$
\begin{aligned}
||\mathbf{v}^T||^2 = ||\mathbf{v}^T||^2 ||\mathbf{v}^*||^2 \geq (\mathbf{v}^T \cdot \mathbf{v}^*)^2 &\geq \left(\gamma \sum_{t=1}^{T} n_t - L\left(\gamma + \sqrt{R^2+1}\right)\right)^2 \\
&\geq \gamma^2 \left(\sum_{t=1}^{T} n^t\right)^2 - 2L\gamma\left(\gamma + \sqrt{R^2+1}\right) \sum_{t=1}^{T} n^t \qquad (5)
\end{aligned}
$$

Crammer と Singer の定理 3 の証明にあるように、

$$
||\mathbf{v}^{T+1}||^2 \leq (R^2(k-1)+1) \sum_{t=1}^{T} n^t \qquad (6)
$$

式 (5) と式 (6) より

$$
\sum_{t=1}^{T} n^t \leq \frac{(k-1)(R^2+1)}{\gamma^2} + 2L\left(1 + \frac{\sqrt{R^2+1}}{\gamma}\right)
$$

が成り立つ。 □

参考文献

[1] S. Ben-David, Nicolo Cesa-Bianchi, D. Haussler and P. M. Long. Characterizations of Learnability for Classes of $\{0, ..., n\}$-Valued Functions. *Journal of Computer and System Sciences* 50, 1995, pp.74-86.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM* 36(4), 1989, pp.929-965.

[3] K. Crammer and Y. Singer. Pranking with Ranking. *Advances in Neural Information Processing* 14, 2002, pp.641-647.

[4] An Introduction to Support Vector Machines. *Cambridge University Press*, 2000.

[5] B. K. Natarajan. On Learning Sets and Functions. *Machine Learning* 4, 1989, pp.67-97.

[6] A. B. Novikoff. On convergence proofs on perceptrons. In Symposium on the Mathematical Theory of Automata 12, 1962, pp.615-622.

[7] S. Rajaram, A. Garg, X. S. Zhou and T. S. Huang. Classification Approach towards Ranking and Sorting Problems. *Proceedings of the 14th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence* 2837, 2003, pp.301-312.